Lu.Getz.Miska_Nature.June.2005.mouse.lung

Module name: Lu.Getz.Miska_Nature.June.2005.mouse.lung
Description: Normal/tumor classifier and *k*NN prediction of mouse lung samples
Author: Gad Getz (Broad Institute), gp-help@broad.mit.edu

Summary
The following description of the analysis is from the supplementary material (http://www.broad.mit.edu/mpr/publications/projects/microRNA/Supplementary_Notes.pdf) of the paper (1):

Normal/tumor classifier and *k*NN prediction of mouse lung samples

In order to build a classifier of normal samples vs. tumor samples based on the miGCM collection, we first picked tissues that have enough normal and tumor samples (at least 3 in each class). The following list summarizes the tissues for this analysis.

Table: Number of Training Samples Used to Build the Normal/Tumor Classifier

Tissue	Number of Normal	Number of Tumor
Colon	5	10
Kidney	3	5
Prostate	8	6
Uterus	9	10
Lung	4	6
Breast	3	6

*k*NN [11] is a predicting algorithm that learns from a training data set (in this case, the above samples from the miGCM data set) and predicts samples in a test data set (in this case, the mouse lung sample set). A set of markers (features that best distinguishes two classes of samples, in this case, normal vs. tumor) was selected using the training data set. Distances between the samples were measured in the space of the selected markers. Prediction is performed, one test sample at a time, by: (i), identifying the *k* nearest samples (neighbors) of the test sample among the training data set; and (ii) assigning the test sample to the majority class of these *k* samples.

We first selected markers that best differentiate the normal and tumor samples (see Supplementary Methods) out of the 187 features that passed the filter (which was applied on the training set alone). This generated a list of 131 markers that each has a p-value <0.05 after Bonferroni correction; 129/131 markers are over-expressed in normal samples, whereas 2/131 are over-expressed in the tumor samples. The following table lists these markers.

Table: Normal/Tumor Makers Selected On the Training Set

Probe	Description	Bonferroni-corrected p-value	Variance-thresholded t-test score
EAM159	hmr_miR-130a	0	10.984
EAM331	hmr_miR-30e	0	10.756
EAM311	hmr_miR-101	0	10.392
EAM299	hmr_miR-195	0	9.957
EAM314	hmr_miR-126	0	9.498
EAM300	h_miR-197	0	8.762
EAM181	hmr_let-7f	0	8.299
EAM380	r_miR-140*	0	8.238

GenePattern

EAM111	hm_let-7g	0	8.235
EAM381	r_miR-151*	0	8.198
EAM218	hmr_miR-152	0	8.180
EAM183	hmr_let-7i	0	8.098
EAM253	hmr_miR-218	0	8.077
EAM155	hmr_miR-136	0	8.058
EAM192	hmr_miR-126*	0	7.991
EAM222	hm_miR-15a	0	7.970
EAM161	hmr_miR-28	0	7.949
EAM184	hmr_miR-100	0	7.894
EAM271	hmr_miR-30c	0	7.848
EAM270	hmr_miR-30b	0	7.731
EAM303	hm_miR-199a*	0	7.519
EAM121	hmr_miR-99a	0	7.515
EAM392	r_miR-352	0	7.476
EAM255	hmr_miR-22	0	7.465
EAM249	hmr_miR-214	0	7.338
EAM160	hmr_miR-26b	0	7.313
EAM133	hmr_miR-324-5p	0	7.266
EAM238	hm_miR-1	0	7.259
EAM179	hmr_let-7d	0	7.235
EAM339	hmr_miR-99b	0	7.225
EAM185	hmr_miR-103	0	7.047
EAM168	hmr_let-7e	0	7.034
EAM200	hmr_miR-133a	0	6.959
EAM278	hmr_miR-98	0	6.952
EAM333	hmr_miR-32	0	6.951
EAM291	hmr_miR-185	0	6.910
EAM187	hmr_miR-107	0	6.879
EAM263	hmr_miR-26a	0	6.818
EAM261	hmr_miR-23b	0	6.814
EAM371	hmr_miR-342	0	6.743
EAM330	hmr_miR-30a-5p	0	6.717
EAM280	hmr_miR-30a-3p	0	6.662
EAM233	hmr_miR-196a	0	6.630
EAM292	hmr_miR-186	0	6.602
EAM115	hmr_miR-16	0	6.558
EAM272	hmr_miR-30d	0	6.516
EAM367	hmr_miR-338	0	6.428
EAM379	r_miR-129*	0	6.323
EAM193	hmr_miR-125a	0	6.222
EAM273	hmr_miR-33	0	6.209
EAM223	hmr_miR-15b	0	6.148
EAM105	hmr_miR-125b	0	6.111
EAM385	hmr_miR-335	0	6.011
EAM237	hmr_miR-19b	0	5.981
EAM320	hm_miR-189	0	5.938
EAM262	hmr_miR-24	0	5.909

EAM240	hmr_miR-20	0	5.908
EAM260	hmr_miR-23a	0	5.901
EAM297	hmr_miR-193	0	5.856
EAM236	hmr_miR-19a	0	5.789
EAM264	hmr_miR-27b	0	5.780
EAM205	hmr_miR-138	0	5.721
EAM234	hmr_miR-199a	0	5.718
EAM207	hmr_miR-140	0	5.561
EAM217	hmr_miR-150	0	5.531
EAM235	h_miR-199b	0	5.516
EAM190	hr_miR-10b	0	5.511
EAM282	m_miR-199b	0	5.483
EAM335	h_miR-34b	0	5.315
EAM288	m_miR-10b	0	5.291
EAM275	hmr_miR-34a	0	5.287
EAM195	hmr_miR-128b	0	5.253
EAM328	hmr_miR-301	0	5.203
EAM365	hmr_miR-331	0	5.191
EAM131	hmr_miR-92	0	5.155
EAM215	hmr_miR-148b	0	5.091
EAM325	hmr_miR-27a	0	5.090
EAM279	hmr_miR-29c	0	5.025
EAM369	hmr_miR-340	0	4.959
EAM354	m_miR-297	0	4.953
EAM119	hmr_miR-29b	0	4.937
EAM210	hmr_miR-143	0	4.908
EAM361	hmr_miR-326	0	4.790
EAM324	hmr_miR-25	0	4.764
EAM226	hmr_miR-181a	0	4.742
EAM343	mr_miR-151	0	4.740
EAM228	hmr_miR-181c	0	4.675
EAM366	mr_miR-337	0	4.661
EAM349	mr_miR-292-3p	0	4.652
EAM189	hmr_miR-10a	0	4.494
EAM355	mr_miR-298	0	4.446
EAM318	h_miR-17-3p	0	4.324
EAM387	r_miR-343	0	4.140
EAM363	mr_miR-329	0	4.118
EAM268	hmr_miR-29a	0	4.044
EAM175	hmr_miR-320	0	3.875
EAM212	hmr_miR-145	0	3.869
EAM378	mr_miR-7b	0	3.853
EAM281	mr_miR-217	0	3.670
EAM307	m_miR-202	0	3.625
EAM209	hmr_miR-142-5p	0	3.594
EAM163	hmr_miR-142-3p	0	3.545
EAM384	r_miR-333	0	3.410
EAM362	hmr_miR-328	0	3.356

EAM329	hm_miR-302a	0	3.348
EAM368	hmr_miR-339	0	3.007
EAM351	m_miR-293	0	2.852
EAM153	hmr_let-7a	0	2.818
EAM360	mr_miR-325	0	2.753
EAM145	hmr_let-7c	0	2.393
EAM348	mr_miR-291-5p	0	2.092
EAM298	hmr_miR-194	0	2.068
EAM250	h_miR-215	0	1.746
EAM229	hm_miR-182	0.005	-4.074
EAM224	hmr_miR-17-5p	0.005	4.875
EAM341	m_miR-106a	0.005	4.185
EAM242	hmr_miR-204	0.005	3.457
EAM295	hmr_miR-190	0.005	3.186
EAM353	m_miR-295	0.005	2.916
EAM246	h_miR-211	0.005	2.663
EAM248	hmr_miR-213	0.01	3.369
EAM186	h_miR-106a	0.01	4.650
EAM137	hmr_miR-132	0.01	3.388
EAM258	hmr_miR-222	0.015	4.257
EAM230	hmr_miR-183	0.02	-3.977
EAM364	mr_miR-330	0.02	3.982
EAM206	hmr_miR-139	0.02	3.761
EAM327	hmr_miR-299	0.025	2.353
EAM232	hmr_miR-192	0.04	1.065
EAM257	hmr_miR-221	0.04	4.321
EAM216	hm_miR-149	0.04	3.711

These 131 markers were used without modification to predict the 12 mouse lung samples using the *k*-nearest neighbour algorithm. Each mouse sample was predicted separately, using $\log_2$ transformed mouse and human expression data. The tumor/normal phenotype prediction of a mouse sample was based on the majority type of the *k* nearest human samples using the chosen metric in the selected feature space. Since the tumor/normal distinction was observed at the raw miRNA expression levels, we decided to use Euclidean distance to measure the distances between samples. Thus, we performed *k*NN with the Euclidean distance measure and *k*=3, resulting in 100% accuracy. The detailed prediction results are available in Supplementary Table 3. Similar classification results were obtained with other *k*NN parameters, with the exception of one mouse tumor T_MLUNG_5 (3rd column from right in Fig. 3b). This sample was occasionally classified as normal, for example, when using cosine distance measure (*k*=3). It should be pointed out that cosine distance captures less an overall shift in expression levels compared to Euclidean distance. It rather focuses on comparing the relationships among the different miRNAs So it appears that the same miRNA data capture different information with different distance metrics; Pearson correlation captures information about the lineage (as seen in clustering results), and Euclidean distance captures the normal/tumor distinction.

References:

- Lu, Getz, Miska, et al. “MicroRNA Expression Profiles Classify Human Cancers,” Nature 435, 834-838 (9 June 2005)